Whose Thumb is It Anyway?: Classifying Author Personality from Weblog Text

Whose thumb is it anyway? Classifying author personality from weblog text

Jon Oberlander
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
j.oberlander@ed.ac.uk

Scott Nowson
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
s.nowson@ed.ac.uk

Abstract

We report initial results on the relatively novel task of automatic classification of author personality. Using a corpus of personal weblogs, or 'blogs', we investigate the accuracy that can be achieved when classifying authors on four important personality traits. We explore both binary and multiple classification, using differing sets of n-gram features. Results are promising for all four traits examined.

1 Introduction

There is now considerable interest in affective language processing. Work focusses on analysing subjective features of text or speech, such as sentiment, opinion, emotion or point of view (Pang et al., 2002; Turney, 2002; Dave et al., 2003; Liu et al., 2003; Pang and Lee, 2005; Shanahan et al., 2005). Discussing affective computing in general, Picard (1997) notes that phenomena vary in duration, ranging from short-lived feelings, through emotions, to moods, and ultimately to long-lived, slowly-changing personality characteristics.

Within computational linguistics, most work has focussed on sentiment and opinion concerning specific entities or events, and on binary classifications of these. For instance, both Pang and Lee (2002) and Turney (2002) consider the thumbs up/thumbs down decision: is a film review positive or negative? However, Pang and Lee (2005) point out that ranking items or comparing reviews will benefit from finer-grained classifications, over multiple ordered classes: is a film review two- or three- or four-star? And at the same time, some work now considers longer-term affective states. For example, Mishne (2005) aims to classify the primary mood of weblog postings; the study encompasses both fine-grained (but non-ordered) multiple classification (frustrated/loved/etc.) and coarse-grained binary classification (active/passive, positive/negative).

This paper is about the move to finer-grained multiple classifications; and also about weblogs. But it is also about even more persistent affective states; in particular, it focusses on classifying *author personality*. We would argue that ongoing work on sentiment analysis or opinion-mining stands to benefit from progress on personality-classification. The reason is that people vary in personality, and they vary in how they appraise events—and hence, in how strongly they phrase their praise or condemnation. Reiter and Sripada (2004) suggest that lexical choice may sometimes be determined by a writer's idiolect—their personal language preferences. We suggest that while idiolect can be a matter of accident or experience, it may also reflect systematic, personality-based differences. This can help explain why, as Pang and Lee (2005) note, one person's four star review is another's two-star. To put it more bluntly, if you're not a very outgoing sort of person, then your thumbs up might be mistaken for someone else's thumbs down. But how do we distinguish such people? Or, if we spot a thumbs-up review, how can we tell whose thumb it is, anyway?

The paper is structured as follows. It introduces trait theories of personality, notes work to date on personality classification, and raises some questions. It then outlines the weblog corpus and the experiments, which compare classification accuracies for four personality dimensions, seven tasks, and five feature selection policies. We discuss the implications of the results, and related work, and end with suggestions for next steps.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 627–634,
Sydney, July 2006.

2 Background: traits and language

Cattell's pioneering work led to the isolation of 16 primary personality factors, and later work on secondary factors led to Costa and McCrae's five-factor model, closely related to the 'Big Five' models emerging from lexical research (Costa and McCrae, 1992). Each factor gives a continuous dimension for personality scoring. These are: Extraversion; Neuroticism; Openness; Agreeableness; and Conscientiousness (Matthews et al., 2003). Work has also investigated whether scores on these dimensions correlate with language use (Scherer, 1979; Dewaele and Furnham, 1999). Building on the earlier work of Gottschalk and Gleser, Pennebaker and colleagues secured significant results using the Linguistic Inquiry and Word Count text analysis program (Pennebaker et al., 2001). This primarily counts relative frequencies of word-stems in pre-defined semantic and syntactic categories. It shows, for instance, that high Neuroticism scorers use: more first person singular and negative emotion words; and fewer articles and positive emotion words (Pennebaker and King, 1999).

So, can a text classifier trained on such features predict the author personality? We know of only one published study: Argamon et al. (2005) focussed on Extraversion and Neuroticism, dividing Pennebaker and King's (1999) population into the top- and bottom-third scorers on a dimension, and discarding the middle third. For both dimensions, using a restricted feature set, they report binary classification accuracy of around 58%: an 8% absolute improvement over their baseline. Although mood is more malleable, work on it is also relevant (Mishne, 2005). Using a more typical feature set (including n-grams of words and parts-of-speech), the best mood classification accuracy was 66%, for 'confused'. At a coarser grain, moods could be classified with accuracies of 57% (active vs. passive), and 60% (positive vs. negative).

So, Argamon et al. used a restricted feature set for binary classification on two dimensions: Extraversion and Neuroticism. Given this, we now pursue three questions. (1) Can we improve performance on a similar binary classification task? (2) How accurate can classification be on the *other* dimensions? (3) How accurate can multiple—three-way or five-way—classification be?

3 The weblog corpus

3.1 Construction

A corpus of personal weblog ('blog') text has been gathered (Nowson, 2006). Participants were recruited directly via e-mail to suitable candidates, and indirectly by word-of-mouth: many participants wrote about the study in their blogs. Participants were first required to answer sociobiographic and personality questionnaires. The personality instrument has specifically been validated for online completion (Buchanan, 2001). It was derived from the 50-item IPIP implementation of Costa and McCrae's (1992) revised NEO personality inventory; participants rate themselves on 41-items using a 5-point Likert scale. This provides scores for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness.

After completing this stage, participants were requested to submit one month's worth of prior weblog postings. The month was pre-specified so as to reduce the effects of an individual choosing what they considered their 'best' or 'preferred' month. Raw submissions were marked-up using XML so as to automate extraction of the desired text. Text was also marked-up by post type, such as purely personal, commentary reporting of external matters, or direct posting of internet memes such as quizzes. The corpus consisted of 71 participants (47 females, 24 males; average ages 27.8 and 29.4, respectively) and only the text marked as 'personal' from each weblog, approximately 410,000 words. To eliminate undue influence of particularly verbose individuals, the size of each weblog file was truncated at the mean word count plus 2 standard deviations.

3.2 Personality distribution

It might be thought that bloggers are more Extravert than most (because they express themselves in public); or perhaps that they are *less* Extravert (because they keep diaries in the first place). In fact, plotting the Extraversion scores for the corpus authors gives an apparently normal distribution; and the same applies for three other dimensions. However, scores for Openness to experience are not normally distributed. Perhaps bloggers are more Open than average; or perhaps there is response bias. Without a comparison sample of matched non-bloggers, one cannot say, and Openness is not discussed further in this paper.

4 Experiments

We are thus confined to classifying on four personality dimensions. However, a number of other variables remain: different learning algorithms can be employed; authors in the corpus can be grouped in several ways, leading to various classification tasks; and more or less restricted linguistic feature sets can be used as input to the classifier.

4.1 Algorithms

Support Vector Machines (SVM) appear to work well for binary sentiment classification tasks, so Argamon et al. (2003) and Pang and Lee (2005) consider One-vs-All, or All-vs-All, variants on SVM, to permit multiple classifications. Choice of algorithm is *not* our focus, but it remains to be seen whether SVM outperforms Naïve Bayes (NB) for personality classification. Thus, we will use both on the binary Tasks 1 to 3 (defined in section 4.2.1), for each of the personality dimensions, and each of the manually-selected feature sets (Levels I to IV, defined in section 4.3). Whichever performs better overall is then reported in full, and used for the multiple Tasks 4 to 7 (defined in section 4.2.2). Both approaches are applied as implemented in the WEKA toolkit (Witten and Frank, 1999) and use 10-fold cross validation.

4.2 Tasks

For any blog, we have available the scores, on continuous scales, of its author on four personality dimensions. But for the classifier, the task can be made more or less easy, by grouping authors on each of the dimensions. The simplest tasks are, of course, binary: given the sequence of words from a blog, the classifier simply has to decide whether the author is (for instance) high or low in Agreeableness. Binary tasks vary in difficulty, depending on whether authors scoring in the middle of a dimension are left out, or not; and if they are left out, what proportion of authors are left out.

More complex tasks will also vary in difficulty depending on who is left out. But in the cases considered here, middle authors are now included. For a three-way task, the classifier must decide if an author is high, medium or low; and those authors known to score between these categories may, or may not, be left out. In the most challenging five-way task, no-one is left out. The point of considering such tasks is to gradually approximate the most challenging task of all: continuous rating.

4.2.1 Binary classification tasks

In these task variants, the goal is to classify authors as either high or low scorers on a dimension:

1. The easiest approach is to keep the high and low groups as far apart as possible: high scorers (H) are those whose scores fall above 1 SD above the mean; low scorers (L) are those whose scores fall below 1 SD below the mean.

2. Task-1 creates distinct groups, at the price of excluding over 50% of the corpus from the analysis. To include more of the corpus, parameters are relaxed: the high group (HH) includes anyone whose score is above .5 SD above the mean; the low group (LL) is similarly placed below.

3. The most obvious task (but not the easiest) arises from dividing the corpus in half about the mean score. This creates high (HHH) and low (LLL) groups, covering the entire population. Inevitably, some HHH scorers will actually have scores much closer to those of LLL scorers than to other HHH scorers.

These sub-groups are tabulated in Table 1, giving the size of each group within each trait. Note that in Task-2, the standard-deviation-based divisions contain very nearly the top third and bottom third of the population for each dimension. Hence, Task-2 is closest in proportion to the division by thirds used in Argamon et al. (2005).

	Lowest	...	Highest
1	L	–	H
2	LL	–	HH
3	LLL		HHH
N1	12	–	13
N2	25	–	22
N3	39		32
E1	11	–	12
E2	23	–	24
E3	32		39
A1	11	–	13
A2	22	–	21
A3	34		37
C1	11	–	14
C2	17	–	27
C3	30		41

Table 1: Binary task groups: division method and author numbers. N = Neuroticism; E = Extraversion; A = Agreeableness; C = Conscientiousness.

4.2.2 Multiple classification tasks

4. Takes the greatest distinction between high (H) and low (L) groups from Task-1, and adds a medium group, but attempts to reduce the possibility of inter-group confusion by including only the smaller medium (m) group omitted from Task-2. Not all subjects are therefore included in this analysis. Since the three groups to be classified are completely distinct, this should be the easiest of the four multiple-class tasks.

5. Following Task-4, this uses the most distinct high (H) and low (L) groups, but now considers all remaining subjects medium (M).

6. Following Task-2, this uses the larger high (hH) and low (Ll) groups, with all those in between forming the medium (m) group.

7. Using the distinction between the high and low groups of Task-5 and -6, this creates a 5-way split: highest (H), relatively high (h), medium (m), relatively low (l) and lowest (L). With the greatest number of classes, this task is the hardest.

These sub-groups are tabulated in Table 2, giving the size of each group within each trait.

	Lowest		...		Highest
4	L	–	m	–	H
5	L		M		H
6	Ll		m	hH	
7	L	l	m	h	H
N4	12	–	24	–	13
N5	12		46		13
N6	25		24	22	
N7	12	13	24	9	13
E4	11	–	24	–	12
E5	11		48		12
E6	23		24	24	
E7	11	12	24	12	12
A4	11	–	28	–	13
A5	11		47		13
A6	22		28	21	
A7	11	11	28	8	13
C4	11	–	27	–	14
C5	11		46		14
C6	17		27	27	
C7	11	6	27	13	14

Table 2: 3-way/5-way task groups: division method and author numbers. N = Neuroticism; E = Extraversion; A = Agreeableness; C = Conscientiousness.

4.3 Feature selection

There are many possible features that can be used for automatic text classification. These experiments use essentially word-based bi- and trigrams. It should be noted, however, that some generalisations have been made: all proper nouns were identified via CLAWS tagging using the WMatrix tool (Rayson, 2003), and replaced with a single marker (NP1); punctuation was collapsed into a single marker (<*p*>); and additional tags correspond to non-linguistic features of blogs—for instance, <*SOP*> and <*EOP*> were used the mark the start and end of individual blogs posts. Word n-gram approaches provide a large feature space with which to work. But in the general interest of computational tractability, it is useful to reduce the size of the feature set. There are many automatic approaches to feature selection, exploiting, for instance, information gain (Quinlan, 1993). However, 'manual' methods can offer principled ways of both reducing the size of the set and avoiding overfitting. We therefore explore the effect of different levels of restriction on the feature sets, and compare them with automatic feature selection. The levels of restriction are as follows:

I The least restricted feature set consists of the n-grams most commonly occurring within the blog corpus. Therefore, the feature set for each personality dimension is to be drawn from the same pool. The difference lies in the number of features selected: the size of the set will match that of the next level of restriction.

II The next set includes only those n-grams which were distinctive for the two extremes of each personality trait. Only features with a corpus frequency $\geq$5 are included. This allows accurate log-likelihood G^2 statistics to be computed (Rayson, 2003). Distinct collocations are identified via a three way comparison between the H and L groups in Task-1 (see section 4.2.1) and a third, neutral group. This neutral group contains all those individuals who fell in the medium group (M) for *all four traits in the study*; the resulting group was of comparable size to the H and L groups for each trait. Hence, this approach selects features using only a *subset* of the corpus. N-gram software was used to identify and count collocations within a sub-corpus (Banerjee

and Pedersen, 2003). For each feature found, its frequency and relative frequency are calculated. This permits relative frequency ratios and log-likelihood comparisons to be made between High-Low, High-Neutral and Low-Neutral. Only features that prove distinctive for the H or L groups with a significance of $p < .01$ are included in the feature set.

III The next set takes into account the possibility that, for a group used in Level-II, an n-gram may be used relatively frequently, but only because a small number of authors in a group use it very frequently, while others in the same group use it not at all. To enter the Level-III set, an n-gram meeting the Level-II criteria must also be used by at least 50%[1] of the individuals within the subgroup for which it is reported to be distinctive.

IV While Level-III guards against excessive individual influence, it may abstract too far from the fine-grained variation *within* a personality trait. The final manual set therefore includes only those n-grams that meet the Level-II criteria with $p < .001$, meet the Level-III criteria, and also correlate significantly ($p < .05$) with individual personality trait scores.

V Finally, it is possible to allow the n-gram feature set to be selected automatically during training. The set to be selected from is the broadest of the manually filtered sets, those n-grams that meet the Level-II criteria. The approach adopted is to use the defaults within the WEKA toolkit: Best First search with the CfsSubsetEval evaluator (Witten and Frank, 1999).

Thus, a key question is when—if ever—a 'manual' feature selection policy outperforms the automatic selection carried out under Level-V. Levels-II and -III are of particular interest, since they contain features derived from a subset of the corpus. Since different sub-groups are considered for each personality trait, the feature sets which meet the increasingly stringent criteria vary in size. Table 3 contains the size of each of the four manually-determined feature sets for each of the four personality traits. Note again that the number of n-grams selected from the most frequent in the corpus for Level-I matches the size of the set for Level-II. In addition, the features automatically selected are task-dependent, so the Level-V sets vary in size; here, the Table shows the number of features selected for Task-2.

[1]Conservatively rounded down in the case of an odd number of subjects.

	I	II	III	IV	V
N	747	747	169	22	19
E	701	701	167	11	20
A	823	823	237	36	34
C	704	704	197	22	25

Table 3: Number of n-grams per set.

	Low	High
N	*[was that]* *[NP1 <p> NP1]* *[<p> after]* *[is that]*	*[this year]* *[to eat]* *[slowly <p>]* *[and buy]*
E	*[point in]* *[last night <p>]* *[it the]* *[is to]*	*[and he]* *[cool <p>]* *[<p> NP1]* *[to her]*
A	*[thank god]* *[have any]* *[have to]* *[turn up]*	*[this is not]* *[<p> it is]* *[<p> after]* *[not have]*
C	*[a few weeks]* *[case <p>]* *[okay <p>]* *[the game]*	*[by the way]* *[<p> i hope]* *[how i]* *[kind of]*

Table 4: Examples of significant Low and High n-grams from the Level-IV set.

To illustrate the types of n-grams in the feature sets, Table 4 contains four of the most significant n-grams from Level-IV for each personality class.

5 Results

For each of the 60 binary classification tasks (1 to 3), the performance of the two approaches was compared. Naïve Bayes outperformed Support Vector Machines on 41/60, with 14 wins for SVM and 5 draws. With limited space available, we therefore discuss only the results for NB, and use NB for Task-4 to -7. The results for the binary tasks are displayed in Table 5. Those for the multiple tasks are displayed in Table 6. Baseline is the majority classification. The most accurate performance of a feature set for each task is highlighted

Task	Base	Lv.I	Lv.II	Lv.III	Lv.IV	Lv.V
N1	52.0	52.0	*92.0*	84.0	**96.0**	*92.0*
N2	53.2	51.1	63.8	68.1	*83.6*	**85.1**
N3	54.9	54.9	60.6	53.5	*71.9*	**83.1**
E1	52.2	56.5	91.3	*95.7*	87.0	**100.0**
E2	51.1	44.7	*74.5*	72.3	66.0	**93.6**
E3	54.9	50.7	53.5	59.2	*64.8*	**85.9**
A1	54.2	62.5	**100.0**	**100.0**	*95.8*	**100.0**
A2	51.2	60.5	*81.4*	79.1	72.1	**97.7**
A3	52.1	53.5	60.6	*69.0*	66.2	**93.0**
C1	56.0	52.0	**100.0**	**100.0**	84.0	*92.0*
C2	61.2	54.5	77.3	*81.8*	72.7	**93.2**
C3	57.7	54.9	63.4	*71.8*	70.4	**84.5**

Table 5: Naïve Bayes performance on binary tasks. Raw % accuracy for 4 personality dimensions, 3 tasks, and 5 feature selection policies.

Task	Base	Lv.I	Lv.II	Lv.III	Lv.IV	Lv.V
N4	49.0	49.0	*81.6*	65.3	77.6	**85.7**
N5	64.8	60.6	*76.1*	67.6	67.6	**94.4**
N6	35.2	31.0	47.9	46.5	*66.2*	**70.4**
N7	33.8	31.0	**49.3**	38.0	42.3	47.9
E4	51.1	44.7	*74.5*	59.6	53.2	**78.7**
E5	67.6	60.6	*83.1*	67.6	54.9	**90.1**
E6	33.8	23.9	*53.5*	46.5	46.5	**56.3**
E7	33.8	**44.7**	39.4	29.6	38.0	40.8
A4	53.8	51.9	**90.4**	78.8	67.3	80.8
A5	66.2	59.2	*83.1*	**84.5**	74.6	80.3
A6	39.4	31.0	*67.6*	60.6	56.3	**85.9**
A7	39.4	33.8	**69.8**	*60.6*	50.7	47.9
C4	51.9	53.8	**92.3**	65.4	67.3	82.7
C5	64.8	62.0	*74.6*	69.0	62.0	**83.1**
C6	38.0	39.4	*59.2*	*59.2*	50.7	**78.9**
C7	38.0	36.6	**62.0**	45.1	45.1	49.3

Table 6: Naïve Bayes performance on multiple tasks. Raw % accuracy for 4 personality dimensions, 4 tasks, and 5 feature selection policies.

in **bold** while the second most accurate is marked *italic*.

6 Discussion

Let us consider the results as they bear in turn on the three main questions posed earlier: Can we improve on Argamon et al.'s (2005) performance on binary classification for the Extraversion and Neuroticism dimensions? How accurately can we classify on the four personality dimensions? And how does performance on multiple classification compare with that on binary classification?

Before addressing these questions, we note the relatively good performance of NB compared with 'vanilla' SVM on the binary classification tasks. We also note that automatic selection generally outperforms 'manual' selection; however overfitting is very likely when examining just 71 data points. Therefore, we do not discuss the Level-V results further.

6.1 Extraversion and Neuroticism

The first main question relates to the feature sets chosen, because the main issue is whether word n-grams can give reasonable results on the Extraversion and Neuroticism classification tasks. Of the current binary classification tasks, Task-2 is most closely comparable to Argamon et al.'s. Here, the best performance for Extraversion was returned by the 'manual' Level-II feature set, closely followed by Level-III. The accuracy of 74.5% represents a 23.4% absolute improvement over baseline (45.8% relative improvement; we report relative improvement over baseline because baseline accuracies vary between tasks). The best performance for Neuroticism was returned by Level-IV. The accuracy of 83.6% represents a 30.4% absolute improvement over baseline (57.1% relative improvement).

Argamon et al.'s feature set combined insights from computational stylometrics (Koppel et al., 2002; Argamon et al., 2003) and systemic-functional grammar. Their focus on function words and appraisal-related features was intended to provide more general and informative features than the usual n-grams. Now, it is unlikely that weblogs are easier to categorise than the genres studied by Argamon et al. So there are instead at least two reasons for the improvement we report.

First, although we did not use systemic-functional linguistic features, we did test n-grams selected according to more or less strict policies. So, considering the manual policies, it seems that the Level-IV was the best-performing set for Neuroticism. This might be expected, given that Level-IV potentially overfits, allowing features to be derived from the full corpus. However, in spite of this, Level-II pproved best for Extraversion. Secondly, in classifying an individual as high or low on some dimension, Argamon et al. had

(for some of their materials) 500 words from that individual, whereas we had approximately 5000 words. The availability of more words per individual is to likely to help greatly in training. Additionally, a greater volume of text increases the chances that a long term 'property' such as personality will emerge

6.2 Binary classification of all dimensions

The second question concerns the relative ease of classifying the different dimensions. Across each of Task-1 to -3, we find that classification accuracies for Agreeableness and Conscientiousness tend to be higher than those for Extraversion and Neuroticism. In all but two cases, the automatically generated feature set (V) performs best. Putting this to one side, of the manually constructed sets, the unrestricted set (I) performs worst, often below the baseline, while Level-IV is the best for classifying each task of Neuroticism. Overall, II and III are better than IV, although the difference is not large.

As tasks increase in difficulty—as high and low groups become closer together, and the left-out middle shrinks—performance drops. But accuracy is still respectable.

6.3 Beyond binary classification

The final question is about how classification accuracy suffers as the classification task becomes more subtle. As expected, we find that as we add more categories, the tasks are harder: compare the results in the Tables for Task-1, -5 and -7. And, as with the binary tasks, if fewer mid-scoring individuals are left out, the task is typically harder: compare results for Task-4 and 5. It does seem that some personality dimensions respond to task difficulty more robustly than others. For instance, on the hardest task, the best Extraversion classification accuracy is 10.9% absolute over the baseline (32.2% relative), while the best Agreeableness accuracy is 30.4% absolute over the baseline (77.2% relative). It is notable that the feature set which return the best results—bar the automatic set V—tends to be Level-II, excepting for Neuroticism on Task-6, where Level-IV considerably outperforms the other sets.

A supplementary question is how the best classifiers compare with human performance on this task. Mishne (2005) reports that, for general mood classification on weblogs, the accuracy of his automatic classifier is comparable to human performance. There are also general results on human personality classification performance in computer-mediated communication, which suggest that at least some dimensions can be accurately judged even when computer-mediated. Vazire and Gosling (2004) report that for personal websites, relative accuracy of judgment was, in descending order: Openness > Extraversion > Neuroticism > Agreeableness > Conscientiousness. Similarly, Gill et al. (2006) report that for personal e-mail, Extraversion is more accurately judged than Neuroticism. The current study does not have a set of human judgments to report. For now, it is interesting to note that the performance profile for the best classifiers, on the simplest tasks, appears to diverge from the general human profile, instead ranking on raw accuracy: Agreeableness > Conscientiousness > Neuroticism > Extraversion.

7 Conclusion and next steps

This paper has reported the first stages of our investigations into classification of author personality from weblog text. Results are quite promising, and comparable across all four personality traits. It seems that even a small selection of features found to exhibit an empirical relationship with personality traits can be used to generate reasonably accurate classification results. Naturally, there are still many paths to explore. Simple regression analyses are reported in Nowson (2006); however, for classification, a more thorough comparison of different machine learning methodologies is required. A richer set of features besides n-grams should be checked, and we should not ignore the potential effectiveness of unigrams in this task (Pang et al., 2002). A completely new test set can be gathered, so as to further guard against overfitting, and to explore systematically the effects of the amount of training data available for each author. And as just discussed, comparison with human personality classification accuracy is potentially very interesting.

However, it does seem that we are making progress towards being able to deal with a realistic task: if we spot a thumbs-up review in a weblog, we should be able to check other text in that weblog, and tell whose thumb it is; or more accurately, what *kind* of person's thumb it is, anyway. And that in turn should help tell us how high the thumb is really being held.

8 Acknowledgements

We are grateful for the helpful advice of Mirella Lapata, and our three anonymous reviewers. The second author was supported by a studentship from the Economic and Social Research Council.

References

Shlomo Argamon, Marin Saric, and Sterling S. Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of SIGKDD*, pages 475–480.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City.

Tom Buchanan. 2001. Online implementation of an IPIP five factor personality inventory [web page]. http://users.wmin.ac.uk/∼buchant/wwwffi/introduction.html [Accessed 25/10/05].

Paul T. Costa and Robert R. McCrae, 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528. ACM Press.

Jean-Marc Dewaele and Adrian Furnham. 1999. Extraversion: The unloved variable in applied linguistic research. *Language Learning*, 49:509–544.

Alastair J. Gill, Jon Oberlander, and Elizabeth Austin. 2006. Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, 40:497–507.

Moshe Koppel, Shlomo Argamon, and Arat Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 7th International Conference on Intelligent User Interfaces*.

Gerald Matthews, Ian J. Deary, and Martha C. Whiteman. 2003. *Personality Traits*. Cambridge University Press, Cambridge, 2nd edition.

Gilad Mishne. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

Scott Nowson. 2006. *The Language of Weblogs: A study of genre and individual differences*. Ph.D. thesis, University of Edinburgh.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 115–124.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

James W. Pennebaker and Laura King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count 2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, Ma.

J. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Paul Rayson. 2003. *Wmatrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Ehud Reiter and Somayajulu Sripada. 2004. Contextual influences on near-synonym choice. In *Proceedings of the Third International Conference on Natural Language Generation*.

Klaus Scherer. 1979. Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social Markers in Speech*, pages 147–209. Cambridge University Press, Cambridge.

James G. Shanahan, Yan Qu, and Janyce Weibe, editors. 2005. *Computing Attitude and Affect in Text*. Springer, Dordrecht, Netherlands.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unspervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 417–424.

Simine Vazire and Sam D. Gosling. 2004. e-perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, 87:123–132.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.